DOCUMENT RESUME

ED 319 159 EC 230 846

AUTHOR Johnsen, Susan K.
TITLE The Validity and Reliability of Instruments Used in Identifying Gifted and Talented Students.
PUB DATE 89
NOTE 39p.; Paper presented at the Conference of the National Association for Gifted Children (Orlando, FL, October 30-November 1, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Aptitude Tests; Elementary Secondary Education; *Gifted; *Screening Tests; Testing Problems; Test Reliability; *Test Reviews; *Test Validity

ABSTRACT

This paper rates the technical aspects of 41 tests that match characteristics associated with giftedness, using criteria developed in "A Consumer's Guide to Tests in Print" (Hammill, Brown, Bryant, 1989). A summary of validity studies showing these tests' use in identifying gifted youngsters for special programs is also provided. In this sample of 41 tests, only 32% of the tests meet acceptable standards in the technical area and only 22% report any validity studies with gifted samples. Recommendations are made to professionals involved in the screening and placement of gifted children. Includes 54 references. (Author/PB)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED319159

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

# The Validity and Reliability of Instruments Used in Identifying Gifted and Talented Students

Susan K. Johnsen
Baylor University

Running head: VALIDITY AND RELIABILITY OF INSTRUMENTS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Susan Johnsen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE



EC 230846

Abstract

The author rates the technical aspects of 41 tests that match characteristics associated with giftedness using criteria developed by *A Consumer's Guide to Tests in Print* (Hammill, Brown, & Bryant, 1989). A summary of validity studies showing these tests' use in identifying gifted youngsters for special programs is also provided. In this sample of 41 tests, only 32% of the tests meet acceptable standards in the technical area and only 22% of the tests report any validity studies with gifted samples. Recommendations are made to professionals who are involved in the screening and placement of gifted children.

# The Validity and Reliability of Instruments Used in Identifying Gifted and Talented Students

Professionals find the selection of methods and formal instruments to identify gifted and talented students an important yet difficult task. Most would agree that the methods and instruments used in screening should match the defining characteristics of the gifted population (Richert, Alvino, & McDonnel, 1982). However, Hoge (1988) has indicated that more often the nature of "giftedness" is defined by the selection of instruments rather than by a set of characteristics or a formal definition. For example, a formal definition may include general intelligence, certain aptitudes, achievement, and commitment characteristics while the most frequently used identification methods and instruments focus primarily on achievement tests, grades, and teacher nominations. In this case, achievement becomes the actual defining characteristic of "giftedness". The primary task, therefore, is to locate instruments whose content relates to a school system's definition of giftedness. Not only must they locate instruments that match definition characteristics but they must also find the ones that meet appropriate technical criteria for reliability and validity. Which tests might a school system defend to its patrons or even in a court of law? Many surveys and reports are available that identify tests for specific categories of giftedness, populations, ages, and identification stages (Alvino, McDonnel, & Richert, 1981; Richert, Alvino, & McDonnel, 1982; Richert, 1985). Unfortunately none of these reviews rate or otherwise evaluate the technical aspects of the instruments themselves. How reliable and valid are the norm-referenced instruments that are often used in screening gifted youngsters? While standards are presented in the American Psychological Association's *Standards for Educational and Psychological Tests* (1985), the professional must apply these to specific tests. Given limited time and multiple responsibilities often compounded by a lack of psychometric sophistication, professionals often select instruments based on tradition, familiarity, or popularity rather than applying a systematic set of standards. The primary purpose of this research is to provide (1) useful criteria for evaluating norm-

referenced test instruments, (2) ratings of the technical aspects of tests matching characteristics associated with giftedness, and (3) reported validity studies showing these tests' use in identifying youngsters for gifted programs.

## Criteria for Evaluation of Norm-Referenced Instruments

In *A Consumer's Guide to Tests in Print*, Hammill, Brown, and Bryant (1989) developed a set of criteria that are based on principles of measurement related to standardization and normative scores, reliability, and validity. They also describe, but do not evaluate other nontechnical areas such as administration, scoring, age/grades for intended use, examiner characteristics, and format characteristics. This section will focus on the criteria developed to evaluate the technical aspects of norm-referenced tests. Some of the nontechnical aspects related to use and administration will be included in another section.

### Norming and Normative Scores

Norming refers to the process in which a test is administered to a sample of a population in a consistent fashion. While the sample may represent a local school district, a geographic area, or a specific special population, national norms generally are more useful because they are widely applicable (Hammill, Brown, & Bryant, 1989). This national sample should be representative of the most current census data and should be large enough so that the sample may be randomly selected (Brown & Bryant, 1984; Salvia & Ysseldyke, 1988). Salvia and Ysseldyke (1988) also have pointed out that normative data need to be updated since both national demographics and patterns of test performance change over time. For example, if a test was more than fifteen years old, an entire class of students would have graduated from school.

Normative scores are created by transforming raw scores into scores that are normally distributed, i.e., the "bell-shaped" curve. Standard scores are one kind of normative score. These scores describe a student's distance from the mean (M) or average in terms of the standard deviation (SD). They may include stanines (M =

5, SD = 1.96), T-scores (M = 50, SD = 10), Z-scores (M = 0, SD = 1), and deviation quotients (M = 100, SD = 15). Most intelligence scores are reported as deviation quotients. A student who scores 130 on a test that uses deviation quotients is scoring two standard deviations away from the mean or in the top two percent. Since the mean and standard deviation remain the same, standard scores may be added, subtracted, and manipulated statistically. When evaluating multiple data, they are therefore useful in comparing the results among diverse test instruments.

Another kind of normative score is percentile ranks. These scores indicate the relative standing of a student in relation to the standardization group. For example, a student who scores at the 98%ile has obtained a score that is higher than 98% of the sample in the standardization group or lower than 2% of the standardization group. Many tests report scores using percentiles since they apear to be the most easily understood by parents and teachers. However, percentiles may not be added, subtracted, and manipulated statistically on matrices unless they are compared and transformed to standard scores.

Age and grade equivalents are another kind of normative score. These scores are derived by calculating the average raw score made by students of a specific age or grade level. These scores are then plotted on a graph with raw scores on one axis and grade or age on the other axis. Scores may not be included for every age or grade level. Therefore, a line is drawn through the plotted points to indicate what age or grade might correspond to a specific raw score. While age and grade equivalents are popular, they are deceptive. First, no common performance exists for two students who might obtain a 5.0 language arts total on an achievement test. An obtained score of 7.5 does not mean that a student is reading at the seventh grade, fifth month, or that a student is performing similarly to seventh grade students. Second, since rates of development are uneven and since age and grade equivalents are plotted on an ordinal scale; they are not useful in comparing the results of various tests.

When evaluating the norming process and normative scores, Hammill, Brown, and Bryant (1989) chose to examine four areas in the previous discussion:

normative scores, size of the normative group, demographic characteristics of the normative group, and recency of the normative data. Figure 1 shows the rating

---

Insert Figure 1 about here

---

form used in the *Consumer's Guide* (1989) to evaluate these four areas. As you may notice, three categories exist for each. For example, under normative scores, the first criterion, "No normative scores are reported; only age or grade equivalents are reported; or only ratio IQs or scores are available," is an example of the poorest technical quality (-). The second criterion, "Percentile ranks, quartiles, deciles, centiles, or dichotomous standard scores are reported," is an example of an acceptable technical quality (+). While, the third criterion, "Standard scores are reported," is an example of the highest technical quality (++). If the test authors or publisher report only non-normative scores, the test would receive the lowest rating (-). On the other hand, the test would receive an acceptable rating (+) for percentiles and related scores, and a good rating (++) for standard scores. In all cases, the burden of proof lies with the test author(s). They need to report clearly the size and major demographic characteristics of their sample. The reader is referred to the *Consumer's Guide* (1989) for a more thorough discussion of the rationale for each of the selected criteria in the four areas.

## Reliability

Reliability refers to how consistently a test measures specific factors or constructs. If a test is unreliable or unstable, a student may be "gifted" on one day and "ungifted" on another day. The student's degree of "giftedness" may be related to the test's reliability rather than to the student's actual ability or lack of ability.

The degree of confidence that a user may place in a test relates to the amount of error associated with a specific test score. This error may be associated with content sampling (i.e., internal consistency reliability), time sampling (i.e., stability reliability), and interscorer differences (i.e., scorer reliability) (Anastasi, 1982). Content sampling relates to the degree that the test items relate to one another and

relate to the quality or construct being measured. The more interrelated the items are, the smaller the test error, and the more likely that the items are measuring a similar quality. In such a case, internal consistency reliability would be quite high. Time sampling relates to the stability of a student's performance over time. In other words, if the same test is administered several times and a student performs similarly (i.e., maintains his/her same rank order), the test probably has adequate stability reliability. Finally, interscorer reliability relates to the way that different scorers rate the same test performance. The more similar the ratings, the higher the interscorer reliability.

One of the many reasons for estimating the reliability of a test is to determine the test's standard error of measurement (SEM). Every test has a certain amount of error. The error is determined from the following formula $SEM = SD\sqrt{1-r}$ in which SD refers to the standard deviation of the test score, and r refers to the test score's reliability coefficient. This SEM value provides examiners with the confidence limits that they need to report the score. For example, if an eight-year-old child scores 120 on a group intelligence test that has a SEM of 3 points for that age child, then 68% of the time, the examiner can be confident that the student's true score would lie somewhere between 117 and 123 (i.e., ± 3 points). If the examiner wanted to increase the confidence level to 95%, two standard errors (or more precisely 1.96 times two SEMs) would need to be added to each side of the obtained score. At this level, the student's true score might lie somewhere between 114 and 126 (i.e., ± 9 points). Finally, if the examiner wanted to increase the confidence level to 99%, approximately three standard errors (or more precisely 2.576 times three SEMs) would need to be added to each side of the obtained score. At this highest level of confidence, the student's true score might lie somewhere between 112 and 128 (i.e., ± 8 points). Reliable tests tend to have small standard errors of measurement.

When evaluating reliability, Hammill, Brown, & Bryant (1989) chose to examine two types of reliability: internal consistency reliability and stability reliability. Figure 2 shows the rating form used in the *Consumer's Guide* (1989)to evaluate these two types of reliability. Again, the first criterion is an example of the

poorest technical quality (-); the second criterion, of an acceptable technical quality(+); and the third criterion, of the highest technical quality (++). If the test

---

Insert Figure 2 about here

---

authors do not report reliability data or if the reliability at most ages is below .80, the test would receive the lowest rating (-). On the other hand, the test would receive an acceptable rating (+) for reliabilities at .80 - .89 at most ages, and a good rating (++) for reliabilities at .90 or above at most ages. In all cases, the test author must adequately report reliability studies--their design, the type of coefficient calculated, and the application of age controls. Again, the reader is referred to the *Consumer's Guide* (1989) for a more thorough discussion of the evaluation of each of the selected criteria in the two areas.

## Validity

Validity refers to how well a test measures what it is supposed to measure. For example, if a test is supposed to measure "giftedness" and discriminate between students who have more or less of the characteristics associated with "giftedness," then students who are in programs for the gifted should score better on this test than students who are not in programs for the gifted.

The American Psychological Association (1985) has suggested that test developers report three types of validity: content, criterion-related, and construct. Content validity refers to how well the test itself samples the domain or content area in question. The degree to which a test has this type of validity is often determined by the thoroughness with which the test is developed. Authors who are knowledgeable regarding the attributes of the construct and who identify items in a statistically sound fashion are more likely to develop a test that meets content validity criteria. Sattler (1988) has suggested that a reviewer pose three kinds of questions regarding a test's content validity. These relate to the degree to which the items resemble the construct or domain being measured, the degree to which the items provide for a full range of responses within the domain, and the degree of

mastery that is being measured. Given a test that purports to measure creative thinking, a reviewer might ask: Does it contain items that require responses associated with creativity? Does it measure the full range of creative thinking behaviors? At what level of mastery is creativity being assessed? Were the items selected using a data-based procedure? If an examiner can respond positively to these four questions, then the test probably has good content validity for the construct of creativity.

Criterion-related validity refers to the relationship between the test score and some criterion measure. Criteria may include future performance (i.e., predictive validity), performance on other tests that measure the same construct (i.e., concurrent validity), or performance by others who share the same attributes or category. If a test for "giftedness" has adequate criterion-related validity, a high score should indicate a student's ability to perform well in classrooms for the gifted, perform similarly on other tests that measure "giftedness," and perform as well as other gifted peers. If a new test of intelligence is being compared with an established one, the established test should have evidence of adequate reliability and validity.

Construct validity refers to the degree that an examiner may make inferences regarding the characteristics or attributes that the test is supposed to be measuring. To study this type of validity, hypotheses are generated regarding the specific construct being measured. Studies are then designed and conducted to determine whether the test is able to support these hypotheses. For example, a test that measures a developmental reasoning ability should correlate with chronological age; a test that measures intelligence should produce different scores for retardates and gifted students; and a test that measures creativity might predict future creative achievements in school or adult life.

While these three types of validity overlap and are not as precisely categorized as reliability, validity studies are necessary in determining the usefulness of a test for identification, screening, and diagnosis (Hammill, Brown, & Bryant, 1989). If a test is to be used to identify gifted students, then, obviously, validity studies should indicate its usefulness in this area. When evaluating validity, Hammill, Brown, & Bryant (1989) chose to examine the various types of research that are

normally conducted in studying test validity. Figure 3 shows the rating form used in the *Consumer's Guide* (1989) to evaluate these types of research.

---

Insert Figure 3 about here

---

A study is acceptable if the technical manual (1) provides sufficient information regarding the research design, (2) reports significant results, and (3) reports results that support the proposed hypotheses. Sufficient information should include the characteristics of the sample, the procedures for data collection, and the nature of data analysis. A study was deemed significant if it reached the .05 level of confidence. In addition, when correlational and intercorrelational coefficients were cited, at least half of these had to reach the stated criteria of .35 in magnitude. Finally, if a positive or negative relationship was supposed to exist between the test and a particular characteristic, or if no relationship at all was hypothesized, then a study should support the relationship.

If the test authors do not report any validity studies or report validity studies in two or fewer categories, the test would receive the lowest rating (-). On the other hand, the test would receive an acceptable rating (+) for reporting validity studies in three to four categories, and a good rating (++) for five or more categories. The reader is referred to the *Consumer's Guide* (1989) for a more thorough discussion of the evaluation of each of the research categories.

## Ratings of Norm-Referenced Instruments

This next section will focus on the evaluation of norm-referenced instruments that cover those categories most often associated with the characteristics of giftedness--general intelligence, aptitudes including creativity and critical thinking, achievement, and non-intellectual areas such as task commitment and personality (Tannenbaum, 1983). This evaluation will focus on the technical aspects related to standardization and norming, reliability, and validity and on the nontechnical characteristics of testing time, group/individual administration, and ages/grades for intended use.

## Selection of Instruments for Review

Instruments were selected based upon the following three criteria: (1) The tests had to be norm-referenced; (2) They had to be relevant to the identification of gifted students; and (3) They had to be available for review. First, instruments that were reviewed needed to have a reference group against whose test performance the scores of other individuals would be compared. While subjective measures may provide clinical data (i.e., how a student goes about acquiring knowledge or developing creative products), such tests do not lend themselves to the same critical review of their technical aspects as a norm-referenced device. Subjective measures are developed for informal use with special populations and and should have demonstrated reliability and validity, but the evaluative criteria applied would be quite different from those used with a norm-referenced test.

Second, instruments were reviewed based upon their relevance to the field of gifted education. Only those tests that purported to measure constructs associated with giftedness were included, including constructs related to intellectual or general ability, aptitudes such as creativity and critical thinking, non-intellectual factors such as personality and task commitment. Since research has indicated that a multiplicity of factors influence the development of giftedness, all of the above areas were included in the review (Richert, Alvino, & McDonnel, 1982).

Third, all reviews were based upon only data provided by the author or publisher, including technical manuals. If the technical data were unavailable or difficult to obtain from the publisher or the manual, these tests were not included in this review. Special attention was given to reliability and/or validity studies that included "gifted" students.

## Review of Test Instruments

Tests were reviewed using the previously discussed criteria established for *A Consumer's Guide to Tests in Print* (Hammill, Brown, & Bryant, 1989). Test reviews extracted from the *Consumer's Guide* are noted with an asterisk (*). All

other tests were reviewed by the author who also surveyed all of the technical manuals to locate any validity studies that included samples of gifted children or youth. If the technical manual simply reported such a study, a summary of the results was listed in Figure 4 under the category, "Studies with Gifted Sample." Unless otherwise indicated, all rated scores are the Battery Composite or Total Scale score associated with a particular test.

---

Insert Figure 4 about here

---

In all, 41 tests were reviewed and are listed alphabetically under the column "Test Name." To examine the application of the criteria, look at the *Cognitive Abilities Test* (CoGat) (Thorndike & Hagen, 1982).

**Nontechnical Areas.** The nontechnical areas related to use and administration are reviewed first. Looking at the first column, this test is placed in the Aptitude category (A) because the content of the test deals with information normally acquired either developmentally or without formal instruction. On the other hand, if the test content had dealt with information acquired through formal instruction, then it would have been coded achievement (Ach); if it had dealt with social or emotional information, its code would have been affective (Aff); and if it had dealt with two or more of these areas--aptitude, achievement, and/or affective--it would have been coded I for general intelligence. Looking at the second column, this test is intended to be used with children in grades three through twelve. The third and fourth columns indicate that the verbal subtest takes approximately ninety or more minutes to administer, and the test may be administered in a large group (G).

**Norms.** Following the brief review of nontechnical areas is a review of the more technical areas related to norms, reliability, and validity. Under "Scores," the *Cognitive Abilities Test*, receives a "good" or "++" rating because it reports standard scores. It also receives a "good" or "++" for "Size" because the normative group contains 100 or more subjects in most of the academic grade levels and has at least 1,000 subjects in the total sample. Under "Demographics," it also receives a "good" or "++" rating because the characteristics of the normative group

correspond to five or more of the following variables: gender, parental education, parental occupation, geographic region, race, ethnicity, intelligence, socioeconomic status, or other relevant variables. Under "Recency," it receives an "acceptable" or "+" because normative data were gathered between the years of 1973 through 1982.

**Reliability.** Under both "internal consistency" and "stability," the CoGat received a "good" or "++" rating because reported reliability coefficients were statistically significant and were .90 or greater for more than half of the age intervals reported. In addition, in reporting the reliability study, the author provided sufficient information regarding sample size, subject characteristics, age controls, statistical analyses, and research design for the reviewers to judge the conclusions derived from the data.

**Validity.** The authors of the CoGat reported at least five validity studies in different areas that were well designed and yielded significant, hypothesized results. However, no studies were reported that used a gifted sample.

In summary, the Verbal Subtest of the *Cognitive Abilities Test* received both good and acceptable ratings across all categories. The test, therefore, meets basic standards and is acceptable from a technical stand point. However, since no studies are reported that demonstrate its valid use in discriminating between gifted and nongifted children, the professional should exercise caution when using the test for these purposes.

## Summary and Conclusions

A variety of tests and methods are used in the process of identifying gifted and talented students (O'Connell, 1985). Although tests often determine the nature of giftedness within a district, test selection should be based on the district's definition as well as its program and student characteristics associated with giftedness. Tests also should meet acceptable psychometric standards. In addition, reported studies should show that the test is valid for the purpose of identifying gifted students. In this way, the district will be able to defend its identification procedure to its patrons and find the most effective match between students and programs.

This study examined the validity and reliability of instruments that are used in identifying gifted students for special programs. Instruments were selected based upon their relevance to the definition of giftedness, their availability for review, and and their use of a norm-referenced method of interpretation. Tests were evaluated using criteria developed for *A Consumer's Guide to Tests in Print* (Hammill, Brown, and Bryant, 1989) to evaluate the areas of standardization and normative scores, reliability, and validity. Nontechnical criteria related to the categories of achievement, aptitude, intelligence, and affect; the range of ages and/or grades included in the standardization sample, the administration time, and whether or not the test was to be administered in a group or individually also were reported. All technical manuals were examined to determine if any validity studies were conducted with gifted samples and if these studies indicated the test's ability to distinguish between gifted and nongifted students.

Of the 41 measures reviewed, the majority of the tests or subtests (61%) fell in the aptitude category (N=25). Less than one-third (27%) of the tests or subtests were classifed as achievement tests (N=11), nine (22%) as intelligence, and only three (7%) as affective measures. Only thirteen measures (32%) met acceptable standards in the technical areas; they received "good" or "acceptable" ratings in all categories on all subtests. Only nine tests (22%) reported any studies using a gifted sample. Of these, only three (7%) had studies that examined whether or not the test was able to distinguish between gifted and nongifted subjects.

While many tests were not reviewed, it appears from this sample that many tests may be used in the field of gifted education even when there is no evidence in the published test manuals for such an application of these instruments. This lack of appropriate instruments for identifying gifted children has been recognized by other researchers as well (Stanley, 1976; Richert, 1985). They have identified problems related to "ceiling" effects, the use of tests to measure aptitudes such as creativity or leadership for which they were not designed, and the underepresentation of minorities in classrooms for the gifted. Along with these problems, professionals who use tests that do not meet acceptable standards or have limited validity data put themselves in a vulnerable position when making decisions

about placement of students into gifted and talented programs, indeed about the placement of any students into any kind of program at all.

More research and evaluation appears to be needed to determine which instruments may be used appropriately in the identification of gifted and talented students. First, professionals in school systems and other users may need to conduct their own validity studies. These studies might relate to the test's ability to distinguish between gifted and nongifted or its ability to predict future performance in special programs. Second, test developers and publishers may want to examine the possibility of developing more instruments for the gifted, of conducting validity studies using gifted samples, and of including gifted youngsters in the early stages of test development, particularly item analysis and selection. In this way, "ceiling" problems and qualitative differences in learning might be addressed. A larger percentage of difficult items could be developed at the upper ranges with content more accurately matching those characteristics attributed to giftedness.

In conclusion, a need continues to exist for reliable and valid measures for identifying gifted and talented students. Those professionals involved in the identification process must exercise caution in the placement of gifted children with possibly unacceptable or inappropriate instruments and continue to conduct research in this area.

References

Aiken, L.R. (1985). Psychological testing and assessment (2nd ed.). Boston, MA: Allyn & Bacon.

Alvino, J.J., McDonnel, R.C., & Richert, E.S. (1981). National survey of identification practices in gifted and talented education. Exceptional Children, 48 (2), 124-132.

American Psychological Association (1985). Standards for Educational and Psychological Tests. Washington, DC: American Psychological Association.

Anastasi, A. (1982). Psychological testing (5th ed.). New York, NY: MacMillan.

Brown, L., & Bryant, B. (1984). A consumer's guide to tests in print: The rating system. Remedial and Special Education, 5 (1), 55-61.

Brown, L., & Coleman, M. (1988). Index of Personality Characteristics. Austin, TX: PRO-ED.

Brown, L., Sherbenou, R.J., & Johnsen, S.K. (1982). Test of Nonverbal Intelligence. Austin, TX: PRO-ED.

Brown, V.L., & McEntire, E. (1984). Test of Mathematical Abilities. Austin, TX: PRO-ED.

Brown, V.L., Hammill, D.D., Wiederholt, J.L. (1986). Test of Reading Comprehension. Austin, TX: PRO-ED.

Connolly, A.J. (1988). KeyMath--Revised. Circle Pines, MN: American Guidance Services.

Dunn, L.M., & Dunn, L. (1981). Peabody Picture Vocabulary Test--Revised. Circle Pines, MN: American Guidance Services.

Ginsburg, H.P., & Baroody, A.J. (1983). Test of Early Mathematics Ability. Austin, TX: PRO-ED.

Gonzalez, E. (1986). Human Figures Drawing Test. Austin, TX: PRO-ED.

Hammill, D. (1985). Detroit Tests of Learning Aptitude-2. Austin, TX: PRO-ED.

Hammill, D., & Bryant, B. (1986). Detroit Tests of Learning Aptitude-P (1986). Austin, TX: PRO-ED.

Hammill, D., Brown, L., & Bryant, B. (1989). A Consumer's Guide to Tests in Print. Austin, TX: PRO-ED.

Hammill, D.D., & Larsen, S.C. (1988). Test of Written Language-2. Austin, TX: PRO-ED.

Harris, J., & Wennerholm, T. (1980). California Achievement Tests, Forms C, D. Monterey, CA: CTB/McGraw-Hill.

Hoge, R. (1988). Issues in the definition and measurement of the giftedness construct. Educational Researcher, 17 (7), 12-16.

Hresko, W.P., Reid, D.K., & Hammill, D.D. (1981). Test of Early Language Development. Austin, TX: PRO-ED.

Johnsen, S.K., & Corn, A.L. (1987). Screening Assessment for Gifted Elementary Students. Austin, TX: PRO-ED.

Johnson, G.O., & Boyd, H.F. (1981). Nonverbal Test of Cognitive Skills. San Antonio, TX: The Psychological Corporation.

Kaufman, A.S., & Kaufman, N.L. (1983). Kaufman Assessment Battery for Children. Circle Pines, MN: American Guidance Services.

Kaufman, A.S., & Kaufman, N.L. (1985). Kaufman Test of Educational Achievement--Brief Form. Circle Pines, MN: American Guidance Services.

Kaufman, A.S., & Kaufman, N.L. (1985). Kaufman Test of Educational Achievement--Comprehensive Form. Circle Pines, MN: American Guidance Services.

Levine, M.N. (1982). The Leiter International Performance Scale Handbook. Los Angeles, CA: Western Psychological Services.

Meeker, M., & Meeker, R. (1981). Structure of Intellect Learning Abilities Test. Los Angeles, CA: Western Psychological Services.

Naglieri, J. (1985). Matrix Analogies Test. San Antonio, TX: The Psychological Corporation.

Newcomer, P.L., & Hammill, D.D. (1988). Test of Language Development-2-Primary. Austin, TX: PRO-ED.

Nunnally, J.S. (1978). Psychometric theory (4th ed.). New York, NY: McGraw-Hill.

O'Connell, P. (1985). The state of the states gifted and talented education. Augusta, ME: Maine Department of Educational and Cultural Services.

Otis, A., & Lennon, R. (1979). Otis Lennon School Ability Test. San Antonio, TX: The Psychological Corporation.

Piers, E.V. (1984). Piers-Harris Children's Self-Concept Scale--Revised Manual. Los Angeles, CA: Western Psychological Services.

Reid, D.K., Hresko, W.P., & Hammill, D.D. (1981). Test of Early Reading Ability. Austin, TX: PRO-ED.

Renzulli, J., & Smith, L. (1978). Learning Styles Inventory. Mansfield Center, CT: Creative Learning Press, Inc.

Richert, E.S. (1985). Identification of gifted students: An update. Roeper Review, 8, 68-72.

Richert, E.S., Alvino, J.J., & McDonnel, R.C. (1982). National report on identification: Assessment and recommendations for comprehensive identification of gifted and talented youth. Sewell, NJ: Educational Information and Resource Center.

Rimm, S.B. (1980). Group Inventory for Finding Creative Talent. Watertown, WI: Educational Assessment Service, Inc.

Salvia, J., & Yseeldyke, J. (1988). Assessment in special and remedial education (3rd ed.). Boston, MA: Houghton Mifflin Company.

Sattler, J. (1988). Assessment of children's intelligence and special abilities (3rd ed.). San Diego: Author.

Slosson, R.L. (1985). Slosson Intelligence Test for Children and Adults--Revised. East Aurora, NY: Slosson Educational Publications, Inc.

Stanford 7 Plus Primary I, Forms EFG (1987). San Antonio, TX: The Psychological Corp.

Stanley, J.C. (1976). The study of mathematically precocious youth. Gifted Child Quarterly, 20, 246-283.

Sullivan, E.T., Clark, W.W., & Tiegs, E.W. (1974). Short Form Test of Academic Aptitude. Monterey, CA: CTB/McGraw-Hill.

Tannenbaum, A. (1983). Gifted children. New York, NY: MacMillan.

Terman, L.M., & Merrill, M.A. (1973). Stanford-Binet Intelligence Scale, 1972 Norms Edition. Boston, MA: Houghton Mifflin Co.

Test of Cognitive Skills (1981). Monterey, CA: CTB/McGraw-Hill.

Thorndike, R., & Hagen, E. (1982). Cognitive Abilities Test, Multilevel. Chicago, IL: Riverside Publishing Company.

Thorndike, R.L., Hagen, E.P., & Sattler, J.M. (1986). Stanford-Binet Intelligence Scale, 4th ed. Chicago, IL: Riverside Publishing Company.

Torrance, E.P. (1974). Torrance Tests of Creative Thinking. Bensenville, IL: Scholastic Testing Service.

Watson, G., & Glaser, E. (1980). Critical Thinking Appraisal. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1967). Wechsler Preschool and Primary Scale of Intelligence. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1974). Wechsler Intelligence Scale for Children-Revised. San Antonio, TX: The Psychological Corporation.

Woodcock, R.W. (1977). Woodcock-Johnson Psycho-Educational Battery. Allen, TX: DLM/Teaching Resources.

Footnotes

1 In *A Consumer's Guide to Tests in Print*, Hammill, Brown, and Bryant used "A," "B," and "F" ratings instead of the "-," "+," and "++" used in this article. Due to the negative connotations often ascribed to letter grades, this author chose to modify these ratings.

## Figure Captions

Figure 1. Evaluation of norms.

Figure 2. Evaluation of reliability.

Figure 3. Evaluation of validity.

Figure 4. Evaluation of tests.

A. Normative Scores

_____ No normative scores are reported; only age or grade equivalents are reported; or only ratio IQs or scores are available.

_____ Percentile ranks, quartiles, deciles, centiles, or dichotomous standard scores are reported.

_____ Standard scores are reported.

B. Size of the Normative Group

_____ The size of the normative group is not specified or does not meet the criteria below.

_____ The normative group contains 75 or more subjects in most one-year age intervals or academic grade levels with which the test is intended to be used. In addition, there are 750 or more subjects in the total sample.

_____ The normative group contains 100 or more subjects in every one-year age interval or academic grade level with which the test is intended to be used. In addition, there are 1,000 or more subjects in the total sample.

C. Demographic Characteristics of the Normative Group

_____ The characteristics of the normative group are not specified or do not meet the criteria below.

_____ The characteristics of the normative group correspond approximately to the known characteristics of the specified population on three or four of the following variables: gender, domicile, parental education, parental occupation, geographic region, race, ethnicity, intelligence, socioeconomic status, or other relevant variables (specify: _____).

_____ The characteristics of the normative group correspond approximately to the known characteristics of the specified population on five or more of the following variables: gender, domicile, parental education, parental occupation, geographic region, race, ethnicity, intelligence, socioeconomic status, or other relevant variable (specify: _____).

D. Recency of Normative Data

_____ 1972 or before.

_____ 1973-1982.

_____ 1983 or after.

Note. From A Consumer's Guide to Tests in Print (p. 132) by D. Hammill, L. Brown, and B. Bryant, 1989, Austin, TX: Pro-Ed. Copyright 1989 by Pro-Ed. Reprinted by permission.

A. Internal Consistency Reliability

_____ is not reported; is below .80 at most ages; or is reported with no apparent controls for the effects of age.

_____ is .80 or above at most ages when reported for two or more age intervals spanning no more than three years each; is .80 - .89 for a single group when evidence is presented to demonstrate that the test is not significantly related to age; or is .80 or above when the effects of age are controlled statistically.

_____ is .90 or above at most ages when reported for two or more age intervals spanning no more than three years each; or is .90 or above for a single group when evidence is presented to demonstrate that the test is no significantly related to age.

B. Stability Reliability

_____ is not reported; is below .80 at most ages; or is reported with no apparent controls for the effects of age.

_____ is .80 or above at most ages when reported for two or more age intervals spanning no more than three years each; is .80 - .89 for a single group when evidence is presented to demonstrate that the test is not significantly related to age; or is .80 or above when the effects of age are controlled statistically.

_____ is .90 or above at most ages when reported for two or more age intervals spanning no more than three years each; or is .90 or above for a single group when evidence is presented to demonstrate that the test is not significantly related to age.

Note. From A Consumer's Guide to Tests in Print (p. 132) by D. Hammill, L. Brown, and B. Bryant, 1989, Austin, TX: Pro-Ed. Copyright 1989 by Pro-Ed. Reprinted by permission.

_____ Validity is not reported; validity studies are not acceptable in design; validity studies do not yield significant results as hypothesized; or validity is supported exclusively by nonempirical evidence of content validity.

_____ Validity is supported by studies that provide appropriate evidence that the test score

- _____ is related as hypothesized to age or grade.
- _____ correlates as hypothesized with measures of achievement.
- _____ correlates as hypothesized with measures of aptitude/developmental abilities or general intelligence.
- _____ correlates as hypothesized with measures of affect.
- _____ produces hypothesized multiple correlation/regression results.
- _____ has hypothesized factor structures.
- _____ predicts appropriately over time.
- _____ distinguishes between specified groups.
- _____ intercorrelates as hypothesized with the test's other scores.
- _____ has appropriate item-total correlations, item discriminating power, or evidence of another appropriate data-based approach to item selection.
- _____ other (specify: _____).

Note. From A Consumer's Guide to Tests in Print (p. 132) by D. Hammill, L. Brown, and B. Bryant, 1989, Austin, TX: Pro-Ed. Copyright 1989 by Pro-Ed. Reprinted by permission.

| Test Name Score Rated (Rated score is Battery Composite or Total Score unless otherwise indicated) | Use | | Admin. | | Norms | | | | Reliability | | Validity | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Category | Ages/Grades | Testing Time | Group/Ind. | Scores | Size | Demographics | Recency | Internal Consistency | Stability | Reported Studies | Studies with Gifted Sample |
| California Achievement Tests, Forms C, D (1980) | Ach | K-12.9 grades | 132+ | G | ++ | ++ | ++ | + | ++ | - | ++ | None |
| Cognitive Abilities Test, Multilevel (1982) Verbal | A | 3-12 grades | 90+ | G | ++ | ++ | ++ | + | ++ | ++ | ++ | None |
| Nonverbal | A | 3-12 grades | 90+ | G | ++ | ++ | ++ | + | ++ | + | ++ | None |
| Quantitative | A | 3-12 grades | 90+ | G | ++ | ++ | ++ | + | ++ | + | ++ | None |
| Critical Thinking Appraisal, Forms A&B (1980) | A | 9 - college | 40+ | G | ++ | ++ | - | + | - | - | - | None |
| *Detroit Tests of Learning Aptitude-2 (1985) General Intelligence | A | 6-17.11 years | 60+ | I | ++ | + | ++ | ++ | ++ | + | + | None |
| Verbal | A | 6-17.11 years | 45 | I | ++ | + | ++ | ++ | ++ | + | + | None |
| Nonverbal | A | 6-17.11 years | 45 | I | ++ | + | ++ | ++ | ++ | + | + | None |
| *Detroit Tests of Learning Aptitude-P (1986) General Intelligence | A | 3.6-9.11 years | 45 | I | ++ | ++ | ++ | ++ | ++ | + | ++ | None |
| Verbal | A | 3.6-9.11 years | 45 | I | ++ | + | ++ | ++ | ++ | + | ++ | None |
| Nonverbal | A | 3.6-9.11 years | 45 | I | ++ | + | ++ | ++ | ++ | + | ++ | None |
| Group Inventory for Finding Creative Talent (1980) | A | K-6 grades | 20+ | G | ++ | - | - | + | + | - | + | Discriminates among gifted/non-gifted |

*Reviewed by A Consumer's Guide to Tests in Print (1989)

++ = Good; + = Acceptable; - = Unacceptable

| Test Name Score Rated (Rated score is the Battery Composite or Total Scale unless otherwise indicated) | Use | | Admin. | | Norms | | | | Reliability | | Validity | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Category | Ages/Grades | Testing Time | Group/Ind. | Scores | Size | Demographics | Recency | Internal Consistency | Stability | Reported Studies | Studies with Gifted Sample |
| *Human Figures Drawing Test (1986) | A | 5.0-10.11 years | 7 | G | ++ | ++ | + | ++ | + | + | + | Discriminates between groups of students identified as intellectually exceptional |
| *Index of Personality Characteristics (1988) | Aff | 8.0-17.11 years | 45 | G | ++ | ++ | ++ | ++ | ++ | - | ++ | None |
| *Kaufman Assessment Battery for Children (1983) Achievement Scale | I | 2.0-12.5 years | 45 | I | ++ | ++ | ++ | + | ++ | ++ | ++ | Correlates with S-B, COGAT (V,Q), gifted case study, included 1.5% in standardization |
| Nonverbal Scale | A | 4.0-12.5 years | 45 | I | ++ | ++ | ++ | + | ++ | + | ++ | Correlates with S-B, COGAT (Q, NV), included 1.5% in standardization |
| Mental Processing | A | 2.0-12.5 years | 45 | I | ++ | ++ | ++ | + | ++ | + | ++ | Correlates with S-B, COGAT (NV), case study, included 1.5% in standardization |
| *Kaufman Test of Educational Achievement - Brief Form (1985) | Ach | 1-12 grades | 22 | I | ++ | - | ++ | ++ | ++ | ++ | ++ | None |
| *Kaufman Test of Educational Achievement - Comp Form (1985) | Ach | 1-12 grades | 60+ | I | ++ | +/++ ** | ++ | ++ | ++ | ++ | ++ | None |
| *Keymath - Revised (1988) | Ach | 5.0-13.11 years | 45 | I | ++ | ++ ** | ++ | ++ | ++ | + | + | None |
| Learning Styles Inventory (1978) | A | 4.0-12.0 | 30 | G | - | - | - | + | - | - | - | None |
| Matrix Analogies Test (1985) | A | 5-17 years | 48 | G | ++ | ++ | + | ++ | ++ | - | ++ | None |
| Nonverbal Test of Cognitive Skills (1981) | A | 6-13 years | 40+ | I | + | - | - | + | - | - | - | None |
| Otis-Lennon School Ability Test, Form R, Primary I (1979) | A | 1 grade | 80 | G | ++ | ++ | ++ | + | ++ | + | ++ | None |

S-B=Stanford-Binet CoGAT=Cognitive Abilities Test
V=Verbal Q=Quantitative NV=Nonverbal
++ = Good; + = Acceptable; - = Unacceptable

*Reviewed by A Consumer's Guide to Tests in Print (1989)
**Fall norms appear first; spring norms follow

| Test Name Score Rated (Rated score is Battery Composite or Total Score unless otherwise indicated) | Use | | Admin. | | Norms | | | | Reliability | | Validity | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Category | Ages/Grades | Testing Time | Group/Ind. | Scores | Size | Demographics | Recency | Internal Consistency | Stability | Reported Studies | Studies with Gifted Sample |
| *Peabody Picture Vocabulary Test-Revised (1981) | A | 2.6-40.11 years | 7 | I | ++ | ++ | ++ | + | + | - | + | None |
| *Piers-Harris Children's Self-Concept Scale - Revised Manual (1984) | Aff | 8.0-18.11 years | 22 | G | ++ | + | - | - | + | - | + | Higher self-concept correlated with analytic high IQ subjects; no significant differences due to age, sex, or grade; correlates to achievement, intelligence, grade point average; self-concept scores were higher for high-achieving females |
| *Screening Assessment for Gifted Elementary Students - Normal Norms (1987) Reasoning | A | 7.0-12.11 years | 10+ | G | ++ | ++ | + | ++ | + | - | ++ | Includes separate sample, discriminates gifted/non-gifted; correlates with achievement, aptitude, and intelligence measures |
| School-Acquired Information | Ach | 7.0-12.11 years | 10+ | G | ++ | ++ | + | ++ | + | - | + | Includes separate sample, discriminates gifted/non-gifted; correlates with achievement, aptitude, and intelligence measures |
| Divergent Production | A | 7.0-12.11 years | 7 | G | ++ | ++ | + | ++ | - | - | - | Includes separate sample, discriminates gifted/non-gifted; correlates with achievement |
| Short Form Test of Academic Aptitude (1974) | A | 1.5-12.9 grades | 12+ | G | ++ | ++ | + | - | + | + | ++ | None |
| *Slosson Intelligence Test - Revised (1985) | I | 2.0-18.0 years | 22 | I | ++ | + | - | + | - | + | - | Correlates with the Wechsler Intelligence Scale for Children - Revised (1974) |
| Stanford 7 Plus Primary I, Forms EFG (1987) | Ach | K-12 grades | | G | ++ | ++ | ++ | ++ | ++ | - | ++ | None |
| Stanford-Binet Intelligence Scale 1972 Norms Edition (1973) | I | 2.0-18.0 years | 45+ | I | ++ | + | - | - | - | - | - | None |

*Reviewed by A Consumer's Guide to Tests in Print (1989)

++ = Good; + = Acceptable; - = Unacceptable

| Test Name Score Rated (Rated score is Battery Composite or Total Score unless otherwise indicated) | Use | | Admin. | | Norms | | | | Reliability | | Validity | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Category | Ages/Grades | Testing Time | Group/Ind. | Scores | Size | Demographics | Recency | Internal Consistency | Stability | Reported Studies | Studies with Gifted |
| *Stanford-Binet Intelligence Scale Fourth Edition (1986) | I | 2.0-23.11 years | 60+ | I | ++ | ++ | ++ | ++ | ++ | ++ | + | Correlates with WISC-R, with S-B (1972), mean score was 123.2 |
| Verbal | I | 2.0-23.11 years | 60+ | I | ++ | ++ | ++ | ++ | ++ | + | + | Correlates with WISC-R, with S-B (1972), mean score was 120.5 |
| Abstract/Visual | I | 2.0-23.11 years | 60+ | I | ++ | ++ | ++ | ++ | ++ | + | + | Mean score was 118.9, correlates with WISC-R performance subtest |
| Quantitative Reasoning | Ach | 2.0-23.11 years | 60+ | I | ++ | ++ | ++ | ++ | ++ | - | + | Correlates with WISC-R, mean score was 120.2 |
| Short-Term Memory | A | 2.0-23.11 years | 60+ | I | ++ | ++ | ++ | ++ | ++ | - | + | Correlates with WISC-R, mean score was 119.7 |
| •Structure of the Intellect Learning Abilities Test (1981) | I | 2-12 grades | 60 | G | ++ | - | - | - | - | - | - | Increased creativity on gifted sample; variations in subtest performance |
| Test of Cognitive Skills (1981) | A | 2-11 grades | 47+ | G | ++ | ++ | ++ | + | + | - | + | None |
| *Test of Early Language Development (1981) | A | 3.0-7.11 grades | 22 | I | ++ | ++ | + | + | + | + | ++ | None |
| *Test of Early Mathematics Ability (1983) | A | 4.0-8.11 years | 22 | I | + | - | - | ++ | ++ | ++ | + | None |
| *Test of Early Reading Ability (1981) | Ach | 3.0-7.11 years | 22 | I | ++ | ++ | + | + | ++ | ++ | ++ | None |
| *Test of Language Development - 2 - Primary (1988) | A | 4.0-8.11 years | 45 | I | ++ | ++ | + | + | ++ | + | + | None |

*Reviewed by A Consumer's Guide to Tests in Print (1989)

++ = Good; + = Acceptable; - = Unacceptable

| Test Name Score Rated (Rated score is the Battery Composite or Total Score unless otherwise indicated) | Use | | Admin. | | Norms | | | | Reliability | | Validity | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Category | Ages/Grades | Testing Time | Group/Ind. | Scores | Size | Demographics | Recency | Internal Consistency | Stability | Reported Studies | Studies with Gifted Sample |
| *Test of Mathematical Abilities (1984) | I | 8.6-18.11 years | 60 | I/G | ++ | + | + | ++ | - | + | ++ | None |
| *Test of Nonverbal Intelligence (1982) | A | 5.0-85.11 years | 7 | I | ++ | + | + | + | + | - | ++ | None |
| *Test of Reading Comprehension - Revised (1986) | Ach | 7.0-17.11 years | 60 | G | ++ | + | + | + | ++ | - | + | None |
| *Test of Written Language-2 (1988) | Ach | 7.6-17.11 years | 60+ | G | ++ | + | + | ++ | ++ | + | + | None |
| *The Leiter International Performance Scale Handbook (1982) | A | 2.0-13.0 years | 60+ | I | - | - | - | - | - | + | + | None |
| *Torrance Tests of Creative Thinking (1974) Verbal fluency | A | K-Adult | 22 | G | ++ | ++ | - | + | - | + | - | Correlates with Rorschach Draw-a-Family, FBRCI, more teacher nominations, predicted humor and fantasy behaviors |
| Verbal originality | A | K-Adult | 22 | G | ++ | ++ | - | + | - | - | - | |
| Figural fluency | A | K-Adult | 22 | G | ++ | ++ | - | + | - | - | - | |
| Figural originality | A | K-Adult | 22 | G | ++ | ++ | - | + | - | - | - | |
| *Wechsler Intelligence Scale for Children - Revised (1974) | I | 6.6-16.6 years | 60+ | I | ++ | ++ | ++ | + | ++ | + | + | None |
| Verbal | I | 6.6-16.6 years | 45 | I | ++ | ++ | ++ | + | ++ | + | + | None |
| Performance | A | 6.6-16.6 years | 45 | I | ++ | ++ | ++ | + | ++ | + | + | None |

*Reviewed by A Consumer's Guide to Tests in Print (1989)  FBRCI = Frenkel-Brunswick Revised California Inventory  ++ = Good; + = Acceptable; - = Unacceptable

| Test Name Score Rated (Rated score is Battery Composite or Total Score unless otherwise indicated) | Use | | Admin. | | Norms | | | | Reliability | | Validity | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Category | Ages/Grades | Testing Time | Group/Ind. | Scores | Size | Demographics | Recency | Internal Consistency | Stability | Reported Studies | Studies with Gifted Sample |
| *Wechsler Preschool and Primary Scale of Intelligence (1967) | I | 4.0-6.5 years | 60 | I | ++ | ++ | ++ | - | ++ | ++ | - | None |
| Verbal | I | 4.0-6.5 years | 45 | I | ++ | ++ | ++ | - | ++ | + | - | None |
| Performance | A | 4.0-6.5 years | 45 | I | ++ | ++ | ++ | - | ++ | + | - | None |
| *Woodcock-Johnson Psycho-Educational Battery-Part One (1978) | I | K-Adult | 60+ | I | + | + | ++ | + | ++ | - | ++ | None |
| Brief Scale | I | K-Adult | 7 | I | + | + | ++ | + | ++ | - | ++ | None |
| Verbal | A | K-Adult | 22 | I | + | + | ++ | + | + | - | + | None |
| Reasoning | A | K-Adult | 22 | I | + | + | ++ | + | + | - | + | None |
| *Woodcock-Johnson Psycho-Educational Battery - Part Two (1978) Reading | Ach | K-Adult | 22 | I | + | + | ++ | + | ++ | - | ++ | None |
| Mathematics | Ach | K-Adult | 7 | I | + | + | ++ | + | ++ | - | ++ | None |
| Written Language | Ach | K-Adult | 7 | I | + | + | ++ | + | ++ | - | ++ | None |
| *Woodcock-Johnson Psycho-Educational Battery - Part Three (1978) Scholastic Interest | Aff | K-Adult | 7 | I | + | + | ++ | + | + | - | + | None |
| Nonscholastic Interest | Aff | K-Adult | 7 | I | + | + | ++ | + | + | - | + | None |

*Reviewed by A Consumer's Guide to Tests in Print (1989)

++ = Good; + = Acceptable; - = Unacceptable